Published online 22 September 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D1095–D1102***
doi:10.1093/nar/gkq811

GreenPhylDB v2.0: comparative and functional genomics in plants

Mathieu Rouard[1,*], Valentin Guignon[1,2], Christelle Aluome[1,2], Marie-Angélique Laporte[2], Gaëtan Droc[2], Christian Walde[1], Christian M. Zmasek[3], Christophe Périn[2] and Matthieu G. Conte[1]

[1]Bioversity International - CfL programme Parc Scientifique Agropolis II, 34397 Montpellier, [2]CIRAD, Department BIOS, UMR DAP - TA40/03, 34398 Montpellier, France and [3]Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

Received June 24, 2010; Revised August 19, 2010; Accepted August 23, 2010

ABSTRACT

GreenPhylDB is a database designed for comparative and functional genomics based on complete genomes. Version 2 now contains sixteen full genomes of members of the *plantae* kingdom, ranging from algae to angiosperms, automatically clustered into gene families. Gene families are manually annotated and then analyzed phylogenetically in order to elucidate orthologous and paralogous relationships. The database offers various lists of gene families including plant, phylum and species specific gene families. For each gene cluster or gene family, easy access to gene composition, protein domains, publications, external links and orthologous gene predictions is provided. Web interfaces have been further developed to improve the navigation through information related to gene families. New analysis tools are also available, such as a gene family ontology browser that facilitates exploration. GreenPhylDB is a component of the South Green Bioinformatics Platform (http://southgreen.cirad.fr/) and is accessible at http://greenphyl.cirad.fr. It enables comparative genomics in a broad taxonomy context to enhance the understanding of evolutionary processes and thus tends to speed up gene discovery.

INTRODUCTION

In recent years, an impressive number of advances in genomics and biotechnologies have emerged, leading to an increase in our knowledge of plant genomes. According to the Genome On-Line Database (GOLD) (1), 233 plant genome sequencing projects have been recorded, mainly due to the advances in high-throughput sequencing technologies. Thus, it is important to provide molecular biologists with a fast and reliable way of applying accumulated genomics knowledge of plant models to plants of agronomic interest. Comparative genomics provides a starting point for understanding the molecular basis of biological diversity among plant species (2) and has a great potential to enhance gene discovery related to economically important traits and plant breeding (3,4). The development of GreenPhylDB v 1.0 (5) was motivated by the sequencing of the *Arabidopsis thaliana* and *Oryza sativa* genomes that paved the way for comparative genomics in plants at the whole genome level. Since then, an objective in molecular biology has been to transfer functional information of genes obtained in one species to another species, thus reducing research time and costs. Orthologous genes (genes that diverged by a speciation event) (6) may have conserved a similar function. Thus analyzing genes between species in order to identify orthologous genes is a reliable strategy for functional annotation (7,8). However, establishing orthology in divergent species is not a trivial exercise. Also, identifying orthologs in plants, in particular in flowering plants, is further complicated by the fact that most plants are paleopolyploids, having experienced extensive gene duplications and evolved faster than animals (9–11). With version 2, we broadened the taxonomy range by adding 14 new genomes representative of major phylum of the *plantae* kingdom including rodophytes (red *algae*) (12), chlorophytes (green *algae*) (13,14), mosses (15), lycophytes and flowering plants with monocotyledons (16–19) and dicotyledons (20–25). This enables a broader view of orthologs shared in plants, and provides a powerful yet simple way to follow function diversification among land plants related to

*To whom correspondence should be addressed. Tel: +33 467 611 302; Fax: +33 467 610 334; Email: m.rouard@cgiar.org

molecular evolution of underlying genes. Thus, GreenPhylDB not only facilitates comparative and functional genomics, but also provides fundamental insights in plant gene family evolution.

METHODOLOGY AND RESULTS

Sequence data and clustering

In GreenPhylDB (v1.0), the clustering was originally performed on the protein-coding genes of the model plants, *O. sativa* and *A. thaliana* using TribeMCL (26). As these two genomes are so far the most stably annotated plant genomes, we decided to enrich gene families classification manually annotated in version 1 (5) and not to perform a new clustering from scratch with larger number of genomes of lower quality. A two-step approach was used to extend the original clustering of plant proteins. Both *O. sativa* and *A. thaliana* clustering was updated with new releases. Then, using a BlastP (27) approach, protein sequences of 14 additional full genomes of the *plantae* kingdom were allocated to existing gene families. The resulting unclassified genes, which for most of them correspond to either genus- or species-specific protein missing in *A. thaliana* and *O. sativa*, were clustered using the same procedure described initially for GreenPhylDB v1.0 (5). (See Supplementary Data 1 and 2 for a description of the full procedure).

Gene family annotation

Gathering high-quality information to support annotation. We used high-quality information sources on protein domains and metabolic pathways including InterPro signatures (28), UniProtKB-SwissProt entries (29) and KEGG pathway entries (30). We considered these databases as they include manually annotated data such as PFAM families and PIRSF homeomorphic gene families. All the protein sequences were scanned with InterProScan. Locus tags correspondences were established or searched with stringent blast parameter (>90% identity) on UniProtKB-SwissProt. Locus tags were made with KEGG using mapping files downloaded on their web site. Pubmed identifiers proposed in UniProtKB-SwissProt annotation were also incorporated. To support annotation and to complement lack of information in these sources, we added sequence annotation from dedicated databases (e.g. TAIR).

Annotation tool. Overall consistency of gene families has been manually checked. For that purpose, to help curators to characterize clusters, we developed an annotation tool that provides a synthetic view of these external data sources, including a summary of the InterPro signatures of the cluster (e.g. number of occurrences, percentage, type and specificity inside the clusters). According to the representation of the protein domains in other gene families, curators may consider the domain specificity (Supplementary Data 3). Focus was given to gene families with multi sources or specific InterPro domains. During curation, registered annotators use collected information to harmonize names, allocate synonyms and decide if the cluster may be considered as a superfamily, family, subfamily or group. Relevant Pubmed identifiers can also be added. Finally a qualitative confidence indicator is given. Curators submit their annotations which are versioned and monitored by an administrator before they appear online.

Annotation confidence. Graphical signs indicate to users the status of the gene cluster annotation process, as well as a confidence level based on the data available for this group of genes. There are five signs indicating the confidence of curation applied: 'High' for very confident curation; 'Normal' for confident curation; 'Unknown' when either there is no information or the information is too scarce to conclude; 'Dubious' and 'Clustering error'. 'High' status is given when several data sources converge to a similar annotation. So, it is likely that even with the availability of new genomes, or new gene annotations, this annotation will remain roughly stable.

Phylogenetic-based analyses and orthology inference

The previous methodology applied to the annotated gene families (31) has been conserved but the phylogenomic pipeline code was overhauled to become more robust and faster, and it was updated with more recent versions of required external programs. The phylogenomic pipeline starts with a multi-fasta file containing sequences of a given gene family. As phylogenetic software is time consuming for large-scale analyses, the pipeline was settled on new high-performance computing facilities, allowing high-throughput analyses. The pipeline is composed of usual steps required for phylogenomics (Figure 1) including:

(i) a filtering procedure based on *E*-value calculated on domain architecture using MEME and MAST software v4.4 (32). A cut-off is automatically calculated for sequences with a too low *E*-value with regards to the median *E*-value of sequences composing the cluster. The filtering procedure also removes alternatively spliced products as well as mis-annotated (e.g. truncated genes) and too highly divergent sequences. Filtered sequences are not excluded from the cluster but they are not taken into account in the next steps to avoid long-branch attraction effect and misleading ortholog identification.
(ii) Multiple alignment using MAFFT v6.240 of the full length protein-coding genes (33).
(iii) Masking of the multiple alignments using AL2CO (34) to optimize the alignment for phylogenetic construction, by removing poorly informative amino acid positions. Then, the number of conserved sites after masking and the ratio of the masked columns on total columns are assessed to check the quality of the masking step.
(iv) Phylogenetic construction using PhyML v3 with 100 bootstraps (35).
(v) Gene rooting and orthologous scoring using RIO implemented in a new version (v2 alpha) of the forester package (36) that produces gene trees at

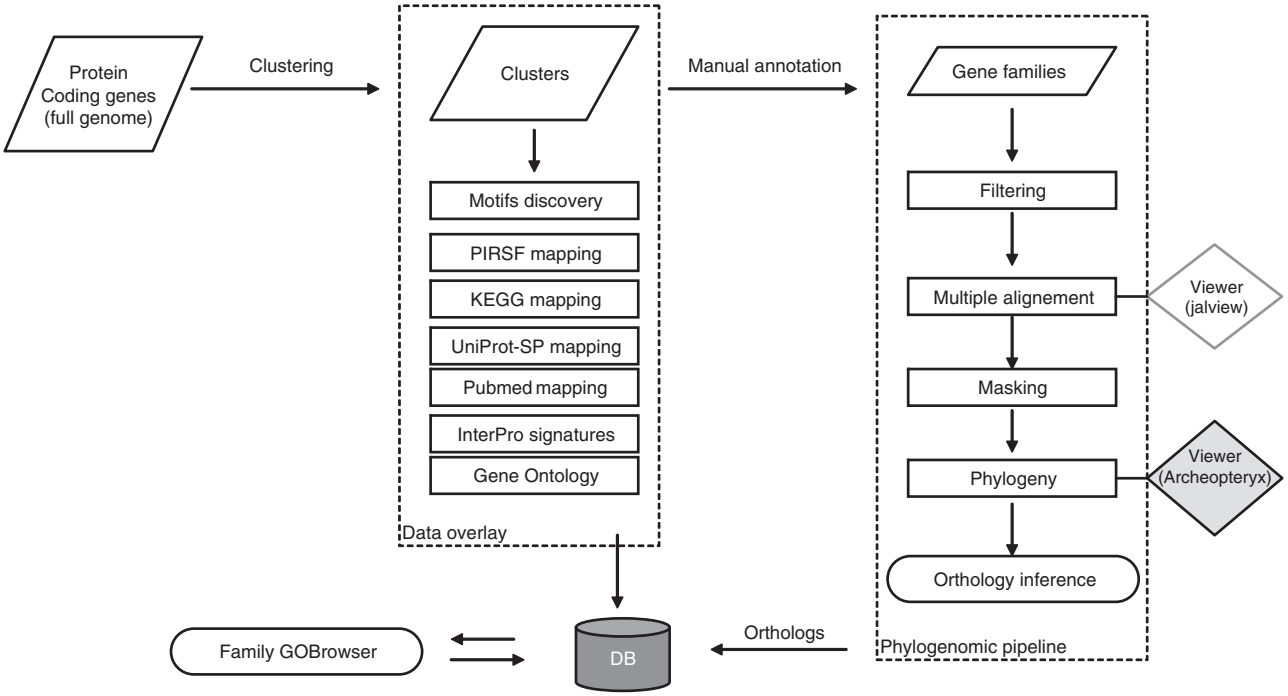

Figure 1. Flowchart of the GreenPhylDB analyses. The input file is a multi-fasta file containing complete plant proteomes. In a first step, an automatic clustering aggregates all proteins in previously defined families. Sequences are classified as orphans if they cannot be regrouped in a cluster. Sequences composing the clusters are analyzed in order to overlay clusters with cross-references (e.g. UniProtKB, Pubmed, InterPro, MEME motifs, KEGG pathways data). Based on this information, clusters are manually curated in order to identify gene families. Finally, gene family sequences are analyzed via a phylogenetic-based pipeline to infer ortholog relationships. The procedure can be iterated for each new released genome using a lighter procedure. This ensures a cumulative and safe growth of the database. The data are stored in the database and can be easily accessed using dedicated visualizing tools including a gene tree viewer, a gene family browser and ortholog extracting tools.

the phyloXML standard (37). Ortholog predictions are based on the concepts (e.g. orthology, sub-tree neighbor, super-orthologs) described by Zmasek and Eddy (36). Bootstrapped rooted trees are used to assign an orthologs score for each sequence.

Using GreenPhylDB

With version 2, the web site has received a face-lift. It now takes advantage of the AJAX technology to increase the speed of user interaction, which is particularly important due to the increasing amount of searchable data within the database. New user-friendly features were developed to facilitate the understanding and interpretation of data. Search facilities or menus on the homepage lead users to lists of gene families or the dedicated webpage for a given gene family or sequence.

The 'Gene family page' is the central page that contains concise information divided in several tabs. A typical gene family overview is shown in Figure 2.

(i) Gene family description: a central table gives a unique identifier, the name and synonyms, some cross-references and the status of the curation and of phylogenetic analyses (Figure 2a).

(ii) Gene family structure: this section allows the user to explore the GreenPhylDB classification, including the number of loci belonging to each cluster (Figure 2b). Sequences were clustered at four levels of stringency, taking into account potential sub-classification (e.g. superfamily, family, subfamily and group). A cluster is often subdivided into different subgroups at a higher stringency level.

(iii) Gene family composition: using the bar chart, users can see the composition of the gene family by species at a glance (Figure 2c). Each bar is clickable and thus produces the list of sequences and their associated cross references (InterPro, KEGG, UniProt, PubMed, GO, etc.).

(iv) Protein domains: in this section, the database provides information about protein domains. Protein-coding genes contained in the database were scanned by InterProscan (38) and Meme suite (Meme and Mast) (32) in order to assess the domain conservation consistency and the specificity of the clustered groups. For each cluster (or gene family), we identified the specificity of InterPro signatures (i.e. found only—or not—in a given cluster) (see Figure 2c and Supplementary Data 3). Meme and Mast analyses may be informative when no InterPro domain exists for a gene family. By sorting on the clustering levels, different profiles are often visible. Even, if biological conclusions cannot easily be reached, it allows assessment of the confidence of the clustering and consideration of those conserved regions not (yet) identified via

Figure 2. Global overview of the family entry page for the Pollen Allergen/Expansin Superfamily (fid = 20923). (**a**) At a glance, users can view that the family is curated (green light) and is plant specific. (**b**) Annotated gene families at the different levels are underlined. Gene families are colored in blue when the phylogenetic analyses are being performed. Here, three gene families are annotated at level 2 (names pop up when you mouse-over) and two of them were analyzed (gene tree and orthologs are available). (**c**) This superfamily contains genes from 15 out of the 16 species. Indeed, there is no representative in the *Cyanidioschyzon merolae*, a red algae and a large expansion is predicted starting in the embryophytes. (**d**) The Expansin/Lol pI InterPro family entry is specific to the Pollen Allergen/Expansin Superfamily. Several other representative domains are listed and graphically represented in a consensus schema. (**e**) Multiple alignment and gene tree Java applets (Jalview and Archeopteryx) including orthology scores can be launched. (**f**) Gene positions on several genomes are available using GViewer. A zoom on chromosome 10 of Zea mays is visible.

interPro. Finally, as it may be informative to graphically view the position of specific protein patterns for a given gene family, we also implemented a consensus pattern graphical view of gene families, based on the Prosite MyDomains Image Creator (39).

(v) Protein list: this section is convenient for displaying a list of genes and their associated sequences. The user can sort associated cross-references for any customized gene list. The final list can be exported in various formats, such as fasta, Excel and CSV.

(vi) Phylogenomics analyses: multiple alignments are available, powered by the latest version of Jalview applet (40), and gene trees can be visualized with the Archeopteryx tree viewer [formerly known as ATV (41)], that was modified to highlight sequence relationships and to display confidence scores for predicted orthologous genes (Figure 2e). The scores in the applet correspond to the orthology score as described in the RIO procedure. An advanced ortholog search bar allows you to display orthologous groups filtered on orthology and Subtree-neighbor scores and distance. Results can be downloaded in various formats such as CSV, XML and Excel.

(vii) Chromosome view: the position of gene family sequences along chromosomes (when the genome has ordered loci) is visible through Flash GViewer (http://gmod.org/wiki/Flashgviewer/) provided by

the GMOD consortium. It acts as a simple synteny viewer, and may help to identify gene duplications or gene clusters (Figure 2f), allowing for visual comparison between plants.

Sequence page

Each sequence stored in GreenPhylDB is accessible through a specific page. Gene identifiers are usually locus tags provided by sequencing consortia, but for some genomes, we prefixed given identifiers by a species code to make them more meaningful (e.g. Phypa_150356 for a sequence of *Physcomitrella patens*). The UniProtKB-SwissProt recommended gene name is also provided if any. To facilitate functional genomic studies, we provided GO annotation and established cross-links with relevant database including microarray (42–45), mutant (46,47) and genome databases (47–49). Links between orthology predictions and gene expressions are suitable for functional studies (Figure 3). Similarly to the gene family page, information is split in tabs containing detailed information around protein sequence, graphical representation of the genomic structure (exons–introns), protein domains and the orthology scores. For the latter, we display all the predicted orthologs for an orthology score >30. It is recommended to also take into consideration subtree-neighbor scores.

Search tools

Search facilities. A keyword search menu is present at the top of every page. In comparison with version 1, auto-completion was added to facilitate searches. Sequence search with BLASTP or BLASTX is available to help users with sequences from other species to extract useful information from GreenPhylDB. In addition, as we produced HMM profiles for all the gene families, users can also use their own sequences to perform a motif structural search using Metameme to compare the domain architecture of the query with the most similar sequence of the database.

Gene family ontology browser. The Gene Ontology (GO) (50) aims at standardizing the representation of gene and gene product attributes across species and databases. It is often used to assign functional annotation to specific genes as implemented with Amigo (51), but large-scale GO term assignments to gene families is also a relevant way to search for a comprehensive set of homologous

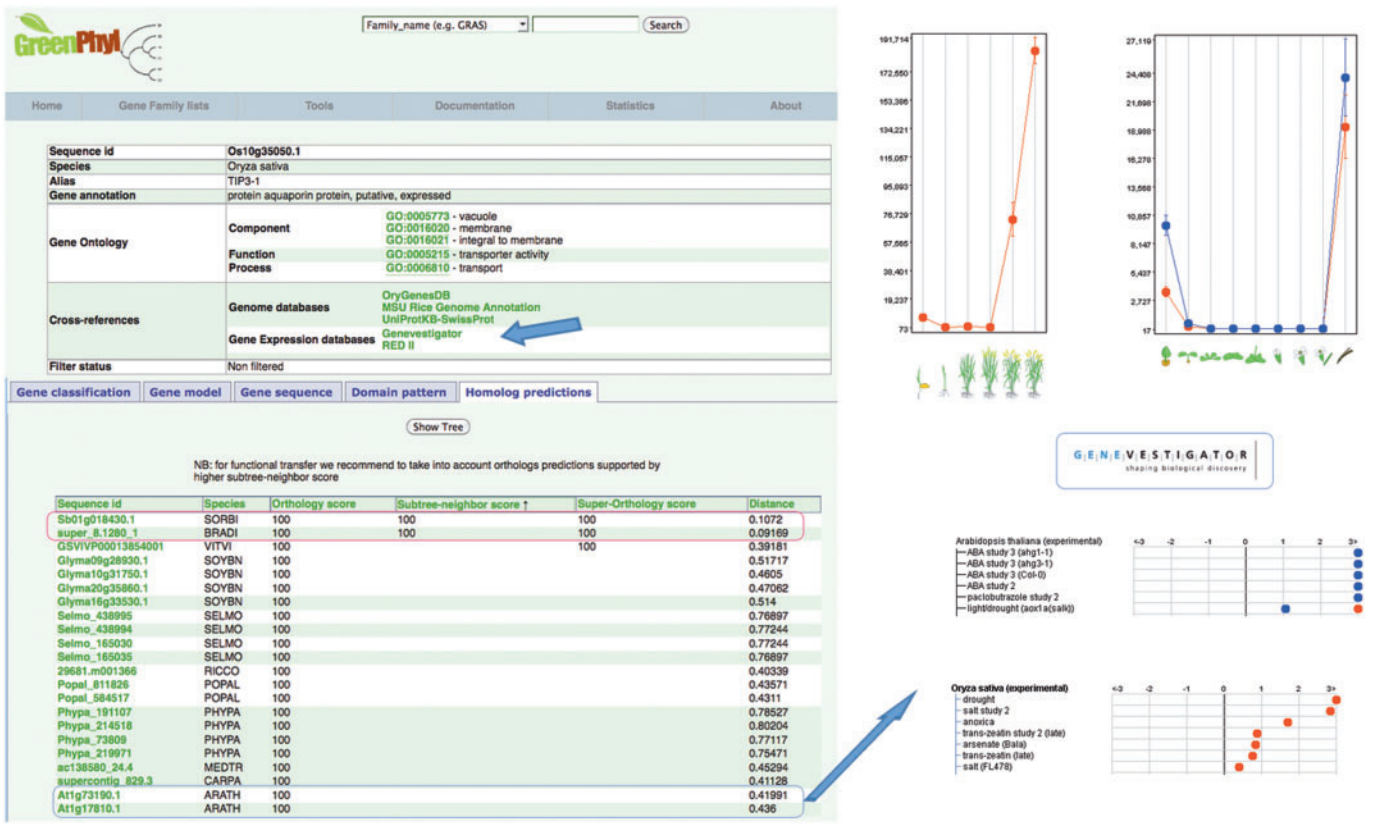

Figure 3. This example illustrates a putative study of a rice gene (Os10g35050.1) and its predicted orthologs in other species of GreenPhylDB. One ortholog gene is found in sorghum (Sb01g018430.1) and in brachypodium (Super_8.1280_1). The query sequence has also two co-orthologs in Arabidopsis (At1g17810.1 in red, At1g73190.1 in blue) that are cross-linked to Genevestigator expression data tools (v3). Os10g35050.1 is over-expressed at the dough stage while At1g17810.1 and At1g73190.1 are expressed in the silique. This may indicate a role in seed development. Moreover, it might be interesting to note that these genes are all over-expressed under drought conditions or in presence of abscissic acid (ABA). This is consistent with the fact that tonoplast-type aquaporins (TIPs) facilitate osmotic water transport across membranes and it suggests a role in response to drought stress.

genes involved in the same process or pathway. Therefore, we developed a light web client to search gene families containing a significant number of genes annotated with the same GO term (http://greenphyl.cirad.fr/cgi-bin/plantslim.cgi) based on GO annotation of InterPro signatures and UniProtKB-SwissProt entries. The web interface displays a digest list of terms defined in the Plant GO (GOslim plants). Users can select a specific GO entry and access a list of families potentially involved in a biological process, a cellular component or a molecular function. More precisely, for each identified family, a user could have access to the sub classification and see if a GO belongs to a specific annotated subgroup.

GreenPhyl database content

To date, GreenPhylDB contains 8231 gene clusters composed of at least five sequences resulting from the clustering of 16 full genome sequences at the clustering level 1. Out of these, 2771 gene families have been annotated with an emphasis on those having specific InterPro protein domains. By default, gene clusters has been tagged as non-curated, however due of the annotation in version 1, some un-annotated clusters may have been assigned a gene family name already. More than 2000 gene families were analyzed with the phylogenomic pipeline. All these numbers will be increasing constantly. Manual annotation will continue in GreenPhyl as new information from high-quality databases increases. Global statistics are provided (http://greenphyl.cirad.fr/v2/cgi-bin/stats.cgi).

Between GreenPhylDB version 1 and this version, we have added a significant number of genomes and made several releases of some of them. Overall, existing clusters remained consistent. Almost all of them were enriched with new sequences. The addition of 14 new genomes modified in some cases orthologous relationships, showing if necessary, that a large-taxonomic sample is suitable for orthologous predictions. We noted an increase of ~25% (~2000) of multiple species gene families. Among new clusters, we noticed are species or phylum specific (e.g. of algae or mosses) but a significant part corresponds to clusters composed of monogenic families (Ex: Photosystem I PsaO Family: 43176). We also noticed that orphan genes (i.e. unclassified sequences) in our classification system were quite often sequences that became obsolete in a later genome release. Multiple-genome clustering is not only useful for comparative genomics, but is also a way to detect low quality sequences in genome annotations.

CONCLUSIONS AND FUTURE DIRECTIONS

The GreenPhyl database has been considerably enriched with additional genomes, and the website has been also overhauled to provide much more user friendly features and tools. It comprises manually annotated gene families with orthology results supported by confidence scores, and connected to key external links such as data expression databases. On a regular basis, we will update the manual annotation of gene families and their phylogenetic analyses. With the increasing numbers of plant sequencing projects, GreenPhylDB will continue to include new genomes, in particular those having a key position in plant taxonomy. Some improvements related to the gene family clustering will be made in order to create seed members. We will also work on new methodologies to propose an orthologous scoring system considering syntenic information.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Vincent Lefort for his support with PhyML and Guilhem Sempere for his modification on the Archeopteryx applet that can now display sequence relationships and its associated scores. We are grateful to Nadège Lanau who contributed to the gene family annotation and to Valentina Barbiero for the homepage species tree. We also thank Manuel Ruiz for his support to the project.

FUNDING

CGIAR Generation Challenge Programme (http://www.generationcp.org) (grant G4008.21). Funding for open access charge: Generation Challenge Programme; National institutes of Health R01 GM087218-01 (to C.M.Z.).

Conflict of interest statement. None declared.

REFERENCES

1. Liolios,K., Chen,I.A., Mavromatis,K., Tavernarakis,N., Hugenholtz,P., Markowitz,V.M. and Kyrpides,N.C. (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **38**, D346–D354.
2. Bowman,J.L., Floyd,S.K. and Sakakibara,K. (2007) Green genes-comparative genomics of the green branch of life. *Cell*, **129**, 229–234.
3. Varshney,R.K., Graner,A. and Sorrells,M.E. (2005) Genomics-assisted breeding for crop improvement. *Trends Plant Sci.*, **10**, 621–630.
4. Flavell,R. (2010) From genomics to crop breeding. *Nat. Biotech.*, **28**, 144–145.
5. Conte,M.G., Gaillard,S., Lanau,N., Rouard,M. and Périn,C. (2008) GreenPhylDB: a database for plant comparative genomics. *Nucleic Acids Res.*, **36**, D991–D998.
6. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
7. Kuzniar,A., van Ham,R.C., Pongor,S. and Leunissen,J.A. (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.*, **24**, 539–551.
8. Gabaldon,T., Dessimoz,C., Huxley-Jones,J., Vilella,A., Sonnhammer,E. and Lewis,S. (2009) Joining forces in the quest for orthologs. *Genome Biol.*, **10**, 403.
9. De Bodt,S., Maere,S. and Van de Peer,Y. (2005) Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.*, **20**, 591–597.
10. Van de Peer,Y., Fawcett,J.A., Proost,S., Sterck,L. and Vandepoele,K. (2009) The flowering world: a tale of duplications. *Trends Plant Sci.*, **14**, 680–688.
11. Salse,J., Abrouk,M., Bolot,S., Guilhot,N., Courcelle,E., Faraut,T., Waugh,R., Close,T.J., Messing,J. and Feuillet,C. (2009) Reconstruction of monocotelydoneous proto-chromosomes reveals

faster evolution in plants than in animals. *Proc. Natl Acad. Sci. USA*, **106**, 14908–14913.
12. Matsuzaki,M., Misumi,O., Shin-i,T., Maruyama,S., Takahara,M., Miyagishima,S., Mori,T., Nishida,K., Yagisawa,F., Nishida,K. *et al.* (2004) Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature*, **428**, 653–657.
13. Merchant,S.S., Prochnik,S.E., Vallon,O., Harris,E.H., Karpowicz,S.J., Witman,G.B., Terry,A., Salamov,A., Fritz-Laylin,L.K., Marechal-Drouard,L. *et al.* (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245–250.
14. Palenik,B., Grimwood,J., Aerts,A., Rouzé,P., Salamov,A., Putnam,N., Dupont,C., Jorgensen,R., Derelle,E., Rombauts,S. *et al.* (2007) The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl Acad. Sci. USA*, **104**, 7705–7710.
15. Rensing,S.A., Lang,D., Zimmer,A.D., Terry,A., Salamov,A., Shapiro,H., Nishiyama,T., Perroud,P., Lindquist,E.A., Kamisugi,Y. *et al.* (2008) The physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.
16. Sequencing ProjectInternational Rice Genome (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
17. Vogel,J., Garvin,D., Mockler,T., Schmutz,J., Rokhsar,D., Bevan,M., Barry,K., Lucas,S., Harmon-Smith,M., Lail,K. *et al.* (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, **463**, 763–768.
18. Paterson,A.H., Bowers,J.E., Bruggmann,R., Dubchak,I., Grimwood,J., Gundlach,H., Haberer,G., Hellsten,U., Mitros,T., Poliakov,A. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.
19. Schnable,P.S., Ware,D., Fulton,R.S., Stein,J.C., Wei,F., Pasternak,S., Liang,C., Zhang,J., Fulton,L., Graves,T.A. *et al.* (2009) The B73 Maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
20. Kaul,S., Koo,H., Jenkins,J., Rizzo,M., Rooney,T., Tallon,L., Feldblyum,T., Nierman,W., Benito,M., Lin,X. *et al.* (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.
21. Jaillon,O., Aury,J., Noel,B., Policriti,A., Clepet,C., Casagrande,A., Choisne,N., Aubourg,S., Vitulo,N., Jubin,C. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.
22. Ming,R., Hou,S., Feng,Y., Yu,Q., Dionne-Laporte,A., Saw,J.H., Senin,P., Wang,W., Ly,B.V., Lewis,K.L.T. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature*, **452**, 991–996.
23. Tuskan,G.A., Difazio,S., Jansson,S., Bohlmann,J., Grigoriev,I., Hellsten,U., Putnam,N., Ralph,S., Rombauts,S., Salamov,A. *et al.* (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, **313**, 1596–1604.
24. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
25. Cannon,S.B., Sterck,L., Rombauts,S., Sato,S., Cheung,F., Gouzy,J., Wang,X., Mudge,J., Vasdewani,J., Schiex,T. *et al.* (2006) Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. *Proc. Natl Acad. Sci.*, **103**, 14959–14964.
26. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
27. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
28. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
29. Schneider,M., Bairoch,A., Wu,C.H. and Apweiler,R. (2005) Plant protein annotation in the UniProt knowledgebase. *Plant Physiol.*, **138**, 59–66.
30. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
31. Conte,M., Gaillard,S., Droc,G. and Perin,C. (2008) Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. *BMC Genomics*, **9**, 183.
32. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucliec Acids Res.*, **37**, W202–W208.
33. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
34. Pei,J. and Grishin,N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.
35. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
36. Zmasek,C.M. and Eddy,S.R. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14.
37. Han,M.V. and Zmasek,C.M. (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, **10**, 356.
38. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
39. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J.A. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, **36**, D245–D249.
40. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
41. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
42. Hruz,T., Laule,O., Szabo,G., Wessendorp,F., Bleuler,S., Oertle,L., Widmayer,P., Gruissem,W. and Zimmermann,P. (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinform.*, Article ID: 420747, doi:10.1155/2008/420747.
43. Gagnot,S., Tamby,J., Martin-Magniette,M., Bitton,F., Taconnat,L., Balzergue,S., Aubourg,S., Renou,J., Lecharny,A. and Brunaud,V. (2008) CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Res.*, **36**, D986–D990.
44. Craigon,D.J., James,N., Okyere,J., Higgins,J., Jotham,J. and May,S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res.*, **32**, D575–D577.
45. Yazaki,J., Kishimoto,N., Ishikawa,M. and Kikuchi,S. (2002) Rice Expression Database: the gateway to rice functional genomics. *Trends Plant Sci.*, **7**, 563–564.
46. Alonso,J.M., Stepanova,A.N., Leisse,T.J., Kim,C.J., Chen,H., Shinn,P., Stevenson,D.K., Zimmerman,J., Barajas,P., Cheuk,R. *et al.* (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, **301**, 653–657.
47. Droc,G., Périn,C., Fromentin,S. and Larmande,P. (2009) OryGenesDB 2008 update: database interoperability for functional genomics of rice. *Nucleic Acids Res.*, **37**, D992–D995.
48. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–1014.
49. Lawrence,C.J., Harper,L.C., Schaeffer,M.L., Sen,T.Z., Seigfried,T.E. and Campbell,D.A. (2008) MaizeGDB: the maize model organism database for basic, translational, and applied research. *Int. J. Plant Genom.*, **2008**, Article ID: 496957, doi:10.1155/2008/496957.

50. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. *Gene Ontology Consort. Nat. Genet*., **25**, 25–29.

51. Carbon,S., Ireland,A., Mungall,C.J., Shu,S., Marshall,B. and Lewis,S. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.